Postprint. Warrens, M. J. (2011). Chance-corrected measures for 2×2 tables that coincide with weighted kappa. *British Journal of Mathematical and Statistical Psychology, 64,* 355-365.

http://dx.doi.org/10.1348/2044-8317.002001

Author. Matthijs J. Warrens
Institute of Psychology
Unit Methodology and Statistics
Leiden University
P.O. Box 9555, 2300 RB Leiden
The Netherlands
E-mail: warrens@fsw.leidenuniv.nl

Chance-Corrected Measures for 2 × 2 Tables That Coincide With Weighted Kappa

Matthijs J. Warrens, Leiden University

Abstract. Cohen's kappa is presently a standard tool for the analysis of agreement in a 2×2 reliability study, and weighted kappa is a standard statistic for summarizing a 2 × 2 validity study. The special cases of weighted kappa, for example Cohen's kappa, are chance-corrected measures of association. For various measures of 2 × 2 association it has been observed in the literature that, after correction for chance, they coincide with a special case of weighted kappa. This paper presents the general function, linear in both numerator and denominator, that becomes weighted kappa after correction for chance.

Key words. Measure of 2 × 2 association; Cohen's kappa; Proportion observed agreement.

1 Introduction

Many results of experimental and research studies can be summarized in a 2×2 table (cf. Warrens, 2008a). An example is a reliability study in which two observers judge each a sample of N subjects on the presence/absence of a trait (Fleiss, 1975; Bloch and Kraemer, 1989). The judgments can be summarized by four counts, namely A, the number of times the observers agreed on the presence of the trait, B, the number of times a trait was present according to the first observer but absent according to the second observer, C, the number of times a trait was absent according to the first observer but present according to the second observer, and D, the number of times the observers agreed on the absence of the trait. A second example is a cluster validation study in which two partitions of a set of data points are compared that were obtained with two different clustering algorithms (Albatineh, Niewiadomska-Bugaj and Mihalko, 2006; Warrens, 2008b). Again the results can be summarized by four counts, namely A, the number of pairs that were placed in the same cluster according to both clustering methods, B (C), the number of pairs that were placed in the same cluster according to the first (second) method but not according to the second (first) method, and D, the number of pairs that were not in the same cluster according to either of the methods.

The four counts that summarize the data can be presented in a 2×2 table. For notational simplicity we will work with the relative frequencies $a = A/N$, $b = B/N$, $c = C/N$ and $d = D/N$ with $a + b + c + d = 1$, instead of the counts A, B, C and D. In general, a 2×2 table can be considered as a cross-classification of two binary $(0, 1)$ variables. A general 2×2 table is presented in Table 1. The row and column totals are the marginal totals that result from summing the relative frequencies. We denote these by p_1 and q_1 for the first variable and by p_2 and q_2 for the second variable.

Insert Table 1 about here.

In many cases a researcher wants to express the degree of association between two binary variables in a single number. The literature contains numerous measures that have been proposed to quantify 2×2 association (Warrens, 2008a,d,e). Many 2×2 measures can be expressed as a function of the quantities a, b, c and d. For example, Cohen's (1960) kappa is presently a standard tool for the analysis of agreement in a 2×2 reliability study, whereas the odds ratio is probably the most widely used measure in epidemiology (Kraemer, 2004).

In a validity study, a binary variable is often compared to a 'gold standard' variable. For example, in a medical test evaluation one has a 'gold standard' evaluation of the presence/absence or type of a disease against which a test is assessed. The weighted kappa index (Spitzer, Cohen, Fleiss

and Endicott, 1967; Bloch and Kraemer, 1989; Kraemer, Periyakoil and Noda, 2004; Warrens, 2010a) is the unique measure that is based on an acknowledgment that the clinical consequences of a false negative may be quite different from the clinical consequences of a false positive. A real number $r \in [0, 1]$ must be specified a priori indicating the relative importance of false negatives to false positives. The sample estimator of the weighted kappa is

$$\kappa(r) = \frac{ad - bc}{rp_1q_2 + (1 - r)p_2q_1}. \tag{1}$$

Bloch and Kraemer (1989) discussed various contexts of association and showed that (1) is the maximum likelihood estimate of the weighted kappa in most cases. The index

$$\kappa\left(\frac{1}{2}\right) = \frac{2(ad - bc)}{p_1q_2 + p_2q_1}$$

is known as Cohen's (1960) kappa (see also, Kraemer, 1979; Warrens, 2008b). Index $\kappa(\frac{1}{2})$ is sometimes called the unweighted kappa.

It should be noted that the measure $\kappa(r)$ is different from the "weighted kappa" coefficient introduced in Cohen (1968) which is discussed in, for example, Fleiss and Cohen (1973), Brenner and Kliebsch (1996), Schuster (2004) and Vanbelle and Albert (2009). The latter descriptive statistic may be applied to square tables larger than 2×2 (Warrens, 2010b). Cohen's (1968) "weighted kappa" is usually applied to cross classifications of two ordinal variables with identical categories and allows the use of weights that quantify the degree of discrepancy between the categories.

The theoretical value of Cohen's kappa and other special cases of weighted kappa is zero when two binary variables are statistically independent. If there is perfect association the maximum value is unity. In various fields of science, zero value under statistical independence is a natural desideratum for an association measure. For example, the desideratum is considered a necessity for reliability studies (Fleiss, 1975) and cluster validation studies (Albatineh et al., 2006; Warrens, 2008b). If a measure of association S does not have zero value under statistical independence, it can be corrected for chance using the linear transformation

$$CS = \frac{S - E(S)}{1 - E(S)} \tag{2}$$

where CS is the chance-corrected measure and $E(S)$ is the value of S expected on the basis of chance alone, calculated from the proportions in Table 2 (Fleiss, 1975; Albatineh et al., 2006; Warrens, 2008a,c). The value 1 in the denominator of (2) represents the maximum value of the association measure S. All 2×2 measures discussed in this paper have a maximum of 1.

Insert Table 2 about here.

For various measures of 2×2 association it has been observed in the literature that, after correction for chance, they coincide with a special case of weighted kappa (Fleiss, 1975; Popping, 1983; Zegers, 1986; Kraemer, 1988; Albatineh et al., 2006; Warrens, 2008c). These findings are gathered in Table 3. For each of the seven measures in Table 3 the numerator and denominator are linear in a, b, c and d. This suggests that there exists a general function of a, b, c and d, of which the measures in Table 3 are special cases, that coincides with the weighted kappa, after correction (2). We derive this particular function in this paper.

Insert Table 3 about here.

The paper is organized as follows. In the next section, we formulate a general function of a, b, c and d that is linear in both numerator and denominator. It is shown which special case of this function is a linear transformation of $a + d$, the observed proportion of agreement, given the marginal totals. Section 3 contains the main result of this paper. In this section it is shown that, after correction for chance, the function derived in Section 2 coincides with weighted kappa. Section 4 contains a discussion.

2 A general function

The literature contains a vast amount of association measures for 2×2 tables (see for example Albatineh et al. (2006) and Warrens (2008a,d,e, 2009) and the references therein). Many of these measures are functions of the relative frequencies a, b, c and d only. Baulieu (1989, 1997) pointed out that many popular functions are fractions with a numerator and denominator that are linear in a, b, c and d. Consider the function

$$\frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}, \tag{3}$$

where weights w_1, w_2, w_3 and w_4 are real numbers. Throughout the paper it is assumed that the value of (3) is defined (Warrens, 2008d). Thus, we avoid cases like $(a, b, c, d) = (0, 0, 0, 0)$, $(w_1, w_2, w_3, w_4) = (0, 0, 0, 0)$, or $(w_1, w_2, w_3, w_4) = \left(\frac{1}{a}, \frac{1}{b}, \frac{-1}{c}, \frac{-1}{d}\right)$. Function (3) is actually over-parametrized because one of the weights can be set to an arbitrary value. However, with four weights instead of three, the formulas that are derived from (3) are symmetric. This leads to a clearer presentation in the remainder of this section.

Function (3) is a rational function with a numerator and denominator that are linear in a, b, c and d. The measures in Table 3 and various

parameter families studied in Tversky (1977), Gower and Legendre (1986) and Baulieu (1989, 1997) are special cases of (3). If w_1, w_2, w_3 and w_4 are nonnegative real numbers, then (3) is increasing in a and d and decreasing in b and c. This property characterizes many measures of 2×2 association (Baulieu, 1989). The maximum value of (3) is 1, which is obtained if $b = c = 0$.

Warrens (2008a,c) studied correction (2) for association measures of a particular form. Consider measures of the form $\lambda + \mu(a + d)$, where λ and μ, unique for each measure, depend on the marginal totals p_1, p_2, q_1 and q_2 of Table 1. Since $a + d$ equals the proportion of observed agreement (see Table 3), all functions of the form $\lambda + \mu(a + d)$ are linear transformations of the observed agreement, given the marginal totals. If we assume that the marginal totals are fixed, then one relative frequency in the 2×2 table determines the rest.

Measures of the form $\lambda+\mu(a+d)$ have been given a lot of attention in the literature (Warrens, 2008a,b). A similar family was studied in Albatineh et al. (2006). These authors considered a family of cluster validation measures of the form $\alpha + \beta \sum_i \sum_j m_{ij}^2$, where m_{ij} is the number of data points that are in cluster i according to the first clustering method and in cluster j according to the second clustering method. We consider two examples and a counterexample of measures that can be expressed in the form $\lambda+\mu(a+d)$.

Example 1. We have

$$a - p_1p_2 = a - (a + b)(a + c) = a(1 - a - b - c) - bc = ad - bc \quad (4)$$

and

$$d - q_1q_2 = d - (b + d)(c + d) = d(1 - b - c - d) - bc = ad - bc. \quad (5)$$

Hence, $a + d - p_1p_2 - q_1q_2 = 2(ad - bc)$, and $\kappa(r)$ can be expressed in the form $\lambda + \mu(a + d)$, where

$$\lambda = \frac{-(p_1p_2 + q_1q_2)}{2rp_1q_2 + 2(1 - r)p_2q_1} \quad \text{and} \quad \mu = \frac{1}{2rp_1q_2 + 2(1 - r)p_2q_1}.$$

Example 2. Since $a = p_2 - q_1 + d$, relative frequencies a and d are linear in $(a + d)$, and $(a + d)$ is linear in a and linear in d. Linear in $(a + d)$ is therefore equivalent to linear in a and linear in d. Hence, each of the seven measures in Table 3 can be expressed in the form $\lambda+\mu(a+d)$. For example, we have

$$\frac{2a}{p_1 + p_2} = \frac{(a + d) - 1}{p_1 + p_2} + 1.$$

The Dice (1945) coefficient can be written in the form $\lambda + \mu(a + d)$ where

$$\lambda = \frac{-1}{p_1 + p_2} + 1 \quad \text{and} \quad \mu = \frac{1}{p_1 + p_2}.$$

Example 3. The Jaccard (1912) coefficient

$$\frac{a}{a+b+c} = \frac{a}{p_1+p_2-a}$$

can be interpreted as the number of positions in which two binary variables both have a 1, divided by the total number of positions in which one of the variable has a 1. The 2×2 measure cannot be expressed in the form $\lambda + \mu(a+d)$.

In general, function (3) is not of the form $\lambda + \mu(a+d)$. Because Warrens (2008a,c) showed that correction (2) is relatively easy for measures of the form $\lambda + \mu(a+d)$, we are interested in what special case of (3) can be expressed in a form $\lambda+\mu(a+d)$. Note that function (3) can only be expressed in the form $\lambda+\mu(a+d)$ if its denominator is a function of the marginal totals p_1, p_2, q_1 and q_2 only. The latter can only be obtained by adding relative frequency a to b or c, or d to b or c. Thus, the numerator of (3) is a function of the marginal totals p_1, p_2, q_1 and q_2 only if the condition

$$w_1 + w_4 = w_2 + w_3 \tag{6}$$

holds. A more substantial interpretation of (6) is the following. If we interpret A and D as the number of positive and negative matches and B and C as the number of mismatches (false positives and false negatives), then (6) requires that the sum of the weights of the matches equals the sum of the weights of the mismatches.

Theorem 1 shows that (6) is also a sufficient condition for the general function (3) to be of the form $\lambda + \mu(a+d)$.

Theorem 1. *Function* (3) *can be expressed in the form* $\lambda+\mu(a+d)$ *if* (6) *holds.*

Proof: We have

$$w_1a + w_4d = \frac{(w_1+w_4)(a+d)}{2} + \frac{(w_1-w_4)(a-d)}{2}, \tag{7}$$

and

$$w_2b + w_3c = \frac{(w_2+w_3)(b+c)}{2} + \frac{(w_2-w_3)(b-c)}{2}. \tag{8}$$

Using $a-d = p_2-q_1$ in (7), we obtain

$$w_1a + w_4d = \frac{(w_1+w_4)(a+d)}{2} + \frac{(w_1-w_4)(p_2-q_1)}{2}. \tag{9}$$

Furthermore, using (6) and the equalities $b-c = p_1-p_2$ and $b+c = 1-a-d$ in (8), we obtain

$$w_2b + w_3c = \frac{w_1+w_4}{2} - \frac{(w_1+w_4)(a+d)}{2} + \frac{(w_2-w_3)(p_1-p_2)}{2}. \tag{10}$$

Hence, using (9) and (10), function (3) can be expressed in the form $\lambda + \mu(a+d)$, where

$$\lambda = \frac{(w_1 - w_4)(p_2 - q_1)}{w_1 + w_4 + (w_1 - w_4)(p_2 - q_1) + (w_2 - w_3)(p_1 - p_2)} \tag{11}$$

and

$$\mu = \frac{w_1 + w_4}{w_1 + w_4 + (w_1 - w_4)(p_2 - q_1) + (w_2 - w_3)(p_1 - p_2)}. \tag{12}$$

This completes the proof.
□

Thus, a special case of (3) that satisfies (6) can be written in the form $\lambda+\mu(a+d)$. Let us specify the form of these association measures. Without loss of generality we may set $w_1 = w$, $w_2 = r$, $w_3 = 1 - r$ and $w_4 = 1 - w$ in (3), (11) and (12). Note that these choices of w_1, w_2, w_3 and w_4 satisfy (6). We obtain, respectively,

$$\frac{wa + (1-w)d}{wa + rb + (1-r)c + (1-w)d}, \tag{13}$$

$$\lambda = \frac{(2w-1)(p_2 - q_1)}{1 + (2w-1)(p_2 - q_1) + (2r-1)(p_1 - p_2)} \tag{14}$$

and

$$\mu = \frac{1}{1 + (2w-1)(p_2 - q_1) + (2r-1)(p_1 - p_2)}. \tag{15}$$

Note that the weight r in (13), (14) and (15) now corresponds to the weight of the weighted kappa in (1). Function (13) can be expressed in the form $\lambda + \mu(a+d)$, where λ and μ are given in (14) and (15) respectively. The seven 2×2 measures in Table 3 are special cases of (13).

3 Correction for chance

Albatineh et al. (2006, p. 309) showed that correction (2) is relatively simple for measures of a form $\alpha + \beta \sum_i \sum_j m_{ij}^2$. Two measures coincide after correction (2) if they have the same ratio (16) (see also Warrens 2008c, p. 490-491).

Lemma 1 [Albatineh et al., 2006]. *Two measures of a form $\lambda+\mu(a+d)$ coincide after correction* (2) *if they have the same ratio*

$$\frac{1-\lambda}{\mu}. \tag{16}$$

Proof: $E(S) = E[\lambda + \mu(a+d)] = \lambda + \mu E(a+d)$ and consequently the CS becomes

$$\begin{aligned} CS &= \frac{S - E(S)}{1 - E(S)} = \frac{\lambda + \mu(a+d) - \lambda - \mu E(a+d)}{1 - \lambda - \mu E(a+d)} \\ &= \frac{a + d - E(a+d)}{\frac{1-\lambda}{\mu} - E(a+d)}. \end{aligned} \tag{17}$$

□

In the previous section we showed that function (13) is of the form $\lambda + \mu(a+d)$. Next we apply Lemma 1 to the function (13).

Theorem 2 shows that two measures of a form (13) coincide after correction (2) if they have the same weight r. Furthermore, this result holds regardless of the value of w.

Theorem 2. *Two functions of a form* (13) *coincide after correction* (2) *if they have the same weight* r.
Proof: Due to Lemma 1, it must be shown that for function (13), the ratio (16) does not depend on the weight w. Plugging the λ and μ in (14) and (15) in ratio (16) we obtain the ratio

$$\frac{1-\lambda}{\mu} = 1 + (2r-1)(b-c). \tag{18}$$

Note that the ratio (18) does not depend on w. This completes the proof.
□

Thus, for fixed r all special cases of (13) coincide after correction for chance. Note that this result holds for any w, and irrespective of the form of $E(a+d)$ (Warrens, 2008c). If we assume that the data are a product of chance concerning two different frequency distributions, one for each variable, the chance-expected value of $a+d$ is estimated as

$$E(a+d) = p_1 p_2 + q_1 q_2 \tag{19}$$

(Table 2; Cohen, 1960; Fleiss, 1975). The estimate in (19) can be obtained by considering all permutations of the observations of one of the variables, while preserving the order of the observations of the other variable. For each permutation the value of $(a+d)$ can be determined. The arithmetic mean of these values is $p_1 p_2 + q_1 q_2$.

We are now ready to present the main result of this paper. Theorem 3 shows that after correction (2), function (13) becomes the weighted kappa statistic.

Theorem 3. *Consider* (19). *Function* (13) *becomes* $\kappa(r)$ *after correction* (2).

Proof: Using (4) and (5), we have $a+d-E(a+d) = 2(ad-bc)$. Furthermore, since

$$p_1 - p_2 = p_1(1-p_2) - p_2(1-p_1) = p_1q_2 - p_2q_1,$$

we have

$$\begin{aligned} 1 + (2r-1)(p_1-p_2) - E(a+d) & \\ &= 2r(p_1-p_2) + p_2 + (1-p_1) - p_1p_2 - q_1q_2 \\ &= 2r(p_1q_2 - p_2q_1) + p_2(1-p_1) + q_1(1-q_2) \\ &= 2rp_1q_2 - 2rp_2q_1 + 2p_2q_1 \\ &= 2rp_1q_2 + 2(1-r)p_2q_1. \end{aligned}$$

Hence, using (18) and (19) in (17), we obtain $\kappa(r)$.
□

4 Discussion

The weighted kappa coefficient $\kappa(r)$ is a standard statistic for summarizing a 2×2 validity study. For various measures of 2×2 association it has been observed in the literature that, after correction for chance, they coincide with a special case of weighted kappa. In this paper we have formalized these observations. Theorem 3, the main result of this paper, shows that function (13) becomes weighted kappa after correction for chance. Another important result is Theorem 1. Function (3) is the general formula of a 2×2 measure with a numerator that is linear in relative frequencies a and d and a denominator that is linear in a, b, c and d. Theorem 1 shows that function (13) is the only function of the form (3) that is a linear transformation of the observed proportion of agreement, given fixed marginal totals, that is, the only special case of (3) that can be expressed in the form $\lambda+\mu(a+d)$. In other words, we have proved in this paper that all 2×2 measures of the form (3) that are linear transformations of the observed proportion of agreement, given fixed marginal totals, become a special case of weighted kappa after correction for chance.

Function (13) is a two-parameter family with parameters $w, r \in \mathbb{R}$. If we restrict r to the range $[0, 1]$, the parameter is analogous to the weight r of weighted kappa. For a fixed value of r, all functions of the form (13) become the special case of weighted kappa corresponding to the same r, regardless of the value of w. We consider three functions of which all special cases coincide after correction for chance.

Using $r = \frac{1}{2}$ and $u = 2w$ in (13) we obtain

$$\frac{ua + (2-u)d}{ua + b + c + (2-u)d}, \qquad u \in \mathbb{R}. \tag{20}$$

The top three statistics in Table 3 are special cases of (20). All functions of the form (20) become Cohen's (1960) kappa ($\kappa(\frac{1}{2})$), after correction for chance. Although many users of Cohen's kappa are not aware of this assumption, Cohen's kappa gives equal weight to the false positives (b) and the false negatives (c) in a validity study. Note that Cohen's kappa is usually interpreted as the chance-corrected version of the proportion of observed agreement ($a + d$), whereas in fact, $\kappa(\frac{1}{2})$ may be interpreted as the chance-corrected version of all functions of a form (20).

Using $r = 1$ in (13) we obtain

$$\frac{wa + (1 - w)d}{wa + b + (1 - w)d}, \qquad w \in \mathbb{R}. \tag{21}$$

The fourth and seventh statistic in Table 3 are special cases of (21). All functions of the form (21) become $\kappa(1) = (ad - bc)/p_1q_2$, after correction (2). If we use $\kappa(1)$ in a validity study, we ignore false negatives and are only concerned with false positives. $\kappa(1)$ may be interpreted as the chance-corrected version of all 2×2 measures of a form (21).

Using $r = 0$ in (13) we obtain

$$\frac{wa + (1 - w)d}{wa + c + (1 - w)d}, \qquad w \in \mathbb{R}. \tag{22}$$

The fifth and sixth statistic in Table 3 are special cases of (22). All functions of the form (21) become $\kappa(0) = (ad - bc)/p_2q_1$, after correction for chance. If we use $\kappa(0)$ in a validity study, we ignore false positives and are primarily interested in false negatives. $\kappa(0)$ may be interpreted as the chance-corrected version of all 2×2 measures of a form (22).

Table 3 presents several special cases of functions (20), (21) and (22). Although the literature contains numerous measures of 2×2 association, we have not found other coefficients that are special cases of (20), (21) or (22). The results presented here can be generalized a bit by combining them with results presented in Warrens (2008c). For example, it follows from Proposition 2 in Warrens (2008c) and Theorem 3 presented here that the function

$$\frac{wa - rb - (1 - r)c + (1 - w)d}{wa + rb + (1 - r)c + (1 - w)d} \tag{23}$$

also becomes weighted kappa after correction for chance. We obtain function (23) by multiplying (13) by 2, followed by subtracting 1. Table 2 presented in Warrens (2008c) contains three additional 2×2 measures that are special cases of this function.

References

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification, 23,* 301-313.

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification, 6,* 233-246.

Baulieu, F. B. (1997). Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification, 14,* 159-170.

Bloch, D. A., & Kraemer, H. C. (1989). 2×2 Kappa coefficients: Measures of agreement or association. *Biometrics, 45,* 269-287.

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology, 7,* 199-202.

Cicchetti, D. V., & Feinstein, A. R. (1990) High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43,* 551-558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-220.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26,* 297-302.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651-659.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613-619.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3,* 5-48.

Jaccard, P. (1912). The distribution of the flora in the Alpine zone. *The New Phytologist, 11,* 37-50.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44,* 461-472.

Kraemer, H. C. (1988). Assessment of 2×2 associations: Generalization of signal-detection methodology. *The American Statistician, 42,* 37-49.

Kraemer, H. C. (2004). Reconsidering the odds ratio as a measure of 2×2 association in a population. *Statistics in Medicine, 23,* 257-270.

Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2004). Tutorial in biostatistics: Kappa coefficients in medical research. *Statistics in Medicine, 21,* 2109-2129.

Popping, R. (1983). *Overeenstemmingsmaten Voor Nominale Data.* PhD. thesis, Groningen, Rijksuniversiteit Groningen.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement, 64,* 243-253.

Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry, 17,* 83-87.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology, 6,* 157-163.

Warrens, M. J. (2008a). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika, 73,* 777-789.

Warrens, M. J. (2008b). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification, 25,* 177-183.

Warrens, M. J. (2008c). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika, 73,* 487-502.

Warrens, M. J. (2008d). On the indeterminacy of resemblance measures for (presence/absence) data. *Journal of Classification, 25,* 125-136.

Warrens, M. J. (2008e). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification, 25,* 195-208.

Warrens, M. J. (2009). k-Adic similarity coefficients for binary (presence/absence) data. *Journal of Classification, 26,* 227-245.

Warrens, M. J. (2010a). A Kraemer-type rescaling that transforms the odds ratio into the weighted kappa coefficient. *Psychometrika, 75,* 328-330.

Warrens, M.J. (2010b). Inequalities between kappa and kappa-like statistics for $k \times k$ tables. *Psychometrika, 75,* 176-185.

Zegers, F. E. (1986). *A General Family of Association Coefficients.* PhD. thesis, Groningen, Rijksuniversiteit Groningen.

Table 1: Break-down of relative frequencies for two binary $(0, 1)$ variables.

	Variable 2		
Variable 1	1	0	Totals
1	a	b	p_1
0	c	d	q_1
Totals	p_2	q_2	1

Table 2: Chance-expected proportions for a 2×2 table.

	Variable 2		
Variable 1	1	0	Totals
1	p_1p_2	p_1q_2	p_1
0	p_2q_1	q_1q_2	q_1
Totals	p_2	q_2	1

Table 3: Descriptions and definitions of various measures of 2×2 association. The third column specifies the special case of weighted kappa that the measure becomes after correction for chance (2).

Description/source	Definition	After correction for chance
Proportion observed agreement	$a + d$	$\kappa(\frac{1}{2})$
Dice (1945) coefficient	$2a/(2a + b + c)$	$\kappa(\frac{1}{2})$
Cicchetti and Feinstein (1990)	$2d/(b + c + 2d)$	$\kappa(\frac{1}{2})$
Sensitivity	$a/(a + b)$	$\kappa(1)$
Specificity	$d/(c + d)$	$\kappa(0)$
Positive predictive value	$a/(a + c)$	$\kappa(0)$
Negative predictive value	$d/(b + d)$	$\kappa(1)$